How to show a probabilistic model is better

Mithun Chakraborty, Sanmay Das, Allen Lavoie
Washington University in St. Louis
{mithunchakraborty,sanmay,allenlavoie}@wustl.edu

Abstract

We present a simple theoretical framework, and corresponding practical procedures, for comparing probabilistic models on real data in a traditional machine learning setting. This framework is based on the theory of proper scoring rules, but requires only basic algebra and probability theory to understand and verify. The theoretical concepts presented are well-studied, primarily in the statistics literature. The goal of this paper is to advocate their wider adoption for performance evaluation in empirical machine learning.

1 Why probabilistic predictions?

When a model is applied to a situation where uncertainty is inherent (e.g. predicting a biased coin flip, or a user's next click), a probability distribution should be its output. Accurate probability distributions provide more information than point predictions, and are the natural product of Bayesian models. Our goal is not to advocate probabilistic models *per se*, but to show in an accessible way that their output can be evaluated rigorously with no more difficulty than deterministic labelings in classification problems.

2 Comparing models

Where do observations come from? They are based on the state of the world. This state describes the situation in which a model is asked to make a prediction.

$$\sigma \sim S \tag{1}$$

The support of this distribution S over states is likely infinite and uncountable. If we are predicting the weather then a state σ includes a description of physical phenomena that could affect future weather patterns. If we are predicting which ad a user will click on, a state includes factors influencing the user's decision: personality, past history, web page design, and so on. The distribution is entirely theoretical, and need never be described formally.

Based on the state of the world σ, an outcome is observed and recorded. However, the outcome is not necessarily implied deterministically by σ. Rather, there is a distribution over possible outcomes:

$$x \sim f_\sigma \tag{2}$$

This includes the possibility of a degenerate distribution (probability 1 on a single outcome), but does not require it. Uncertainty could stem from true randomness (e.g. quantum noise) or from ignorance (e.g. the model does not know what the user ate for breakfast). The noise distribution f_σ is again entirely theoretical, and need never be described.

Equations (1) and (2) define a generative framework for observations. When scoring the probabilistic predictions of a model, we will typically have a single observation from each of many *different* states of the world $\sigma_1 \dots \sigma_n$ (although states drawn multiple times pose no problem). That is, we have a set X of n observations:

$$X = \{x_{\sigma_1} \dots x_{\sigma_n}\} \tag{3}$$

For convenience, we will assume that these observations are discrete, but a generalization to real-valued observations is possible. Corresponding to each of these observations are predictions from each of the models we are evaluating. For simplicity, we assume two models, g and k.

$$G = \{g_{\sigma_1} \dots g_{\sigma_n}\} \tag{4}$$

$$K = \{k_{\sigma_1} \dots k_{\sigma_n}\} \tag{5}$$

Here g_{σ_1} is the distribution that model g predicts in the state σ_1 where the observed outcome is x_{σ_1}, and likewise for the remainder of the observations and for model k. The theoretical assumption is that $x_{\sigma_i} \sim f_{\sigma_i}$, but the states σ_i need no description for the purposes of model evaluation and we never need to construct f_{σ_i} explicitly. To say that model g is "better" than model k, we would like to conclude that it has a lower divergence from the true distribution f in expectation for some divergence function d:

$$E_{\sigma\sim S}[d(f_\sigma||g_\sigma)] < E_{\sigma\sim S}[d(f_\sigma||h_\sigma)] \tag{6}$$

Since we have a finite number n of samples, we can only determine probabilistically if this inequality holds. Examples of d for which this estimation task is possible using only X, G, and K are squared Euclidean distance $d(p||q) = ||p-q||^2$ and KL-divergence $d(p||q) = \sum_j p_j \ln \frac{p_j}{q_j}$. However, it is not immediately obvious how the truth of the inequality in (6) can be evaluated, even probabilistically, *without access to the true distributions* $f_{\sigma_1}, \dots, f_{\sigma_n}$. However, only simple algebra is required. For KL-divergence, we first approximate the expectation of the log probability assigned by model g (the derivation for k is identical) to the true outcome x, that is:

$$E_{\sigma\sim S}\left[E_{x\sim f_\sigma}[-\ln(g_{\sigma,x})]\right] \tag{7}$$

This expression can be approximated from G and X:

$$-\frac{1}{n}\sum_{i=1}^{n}\ln(g_{\sigma_i,x_{\sigma_i}}) \tag{8}$$

Where $g_{\sigma_i,x_{\sigma_i}}$ is the probability that model g assigned to the true outcome x_{σ_i} (corresponding in the theoretical model to state σ_i). The trick is that (7) is equivalent to expected KL-divergence plus a constant:

$$\begin{aligned}
&= E_{\sigma\sim S}\left[\sum_j f_{\sigma,j}(-\ln g_{\sigma,j})\right] \\
&= E_{\sigma\sim S}\left[\sum_j f_{\sigma,j}(-\ln g_{\sigma,j}+\ln f_{\sigma,j}-\ln f_{\sigma,j})\right] \\
&= E_{\sigma\sim S}\left[\sum_j \left(f_{\sigma,j}\ln\frac{f_{\sigma,j}}{g_{\sigma,j}} - f_{\sigma,j}\ln f_{\sigma,j}\right)\right] \\
&= E_{\sigma\sim S}\left[d_{\text{KL}}(f_\sigma||g_\sigma)+H(f_\sigma)\right]
\end{aligned}$$

Here $H(f) = -\sum_j f_j \ln f_j$ is the Shannon entropy of f. Since $H(f_\sigma)$ is independent of a model's predictions, differences in (7) between models g and k must be due to differences in the expected KL-divergences $E_{\sigma\sim S}[d_{\text{KL}}(f_\sigma||g_\sigma)]$ and $E_{\sigma\sim S}[d_{\text{KL}}(f_\sigma||k_\sigma)]$. The only remaining complication is the finite sample: how can we be sure that observed differences in (8) are due to differences in (7)?

This is a standard statistical task: we have a set of n paired samples $(-\ln g_{\sigma_i,x_{\sigma_i}}, -\ln k_{\sigma_i,x_{\sigma_i}})$ related by the state of the world σ_i for each sample, and want to test whether the expectation of the g samples is significantly less than that of the k samples (meaning g is a better model). A paired t-test or the Wilcoxon signed-rank test (although it tests the median rather than the mean) are reasonable options.

This simple algebraic trick comes out of the theory of proper scoring rules (see Gneiting and Raftery [2007] for a thorough survey). Scoring rules were developed to incentivize true reporting of probabilities by experts: first a report is solicited in the form of a probability distribution q, then an outcome is observed. The expert is paid based on their report and the outcome, according to the scoring rule. A proper scoring rule incentivizes an expert to report truthfully (which is not the case if the expert is paid e.g. q_i for an outcome i, often referred to as the naive scoring rule). Any proper scoring rule has an associated divergence function, which for the logarithmic scoring rule ($\ln q_i$ for an observed outcome i) is KL-divergence. The divergence function associated with the quadratic scoring rule $2q_i - ||q||^2$ is squared Euclidean distance, which can also be derived with only simple algebra:

$$\begin{aligned}
&E_{\sigma\sim S}\left[E_{x\sim f_\sigma}\left[-2g_{\sigma,x}+||g_\sigma||^2\right]\right] \tag{9}\\
=&E_{\sigma\sim S}\left[-2f_\sigma\cdot g_{\sigma,j}+||g_\sigma||^2+||f_\sigma||^2-||f_\sigma||^2\right] \tag{10}\\
=&E_{\sigma\sim S}\left[||f_\sigma-g_\sigma||^2-||f_\sigma||^2\right] \tag{11}\\
& \tag{12}
\end{aligned}$$

As with the logarithmic scoring rule, we get a divergence function $||f_\sigma - g_\sigma||^2$ and a generalized entropy term $||f_\sigma||^2$ which again is independent of the model's predictions.

3 Procedure summary

While theoretically justifying probabilistic model comparisons is slightly tedious, the procedure could not be simpler. To summarize:

- For every held-out observation, score each model's predicted distribution q: $-\ln(q_i)$ for logarithmic, or $-2q_i+||q||^2$ for quadratic given that outcome i is observed

- Perform a (typically paired) statistical test to determine whether the scores for one model are significantly lower than those for the other, lower indicating a better model

When comparing more than two models, perform as many pairwise tests as necessary. The "figure of merit" for a model is its mean score (e.g. (8) for the logarithmic scoring rule), lower implying less divergence from the unobserved true distributions of observations and therefore a better model (modulo noise in the estimate).

4 Choosing a divergence function

We have presented two of the most common choices for scoring rules, quadratic and logarithmic, corresponding to evaluation with squared Euclidean distance and KL-divergence respectively. The former is of interest when KL-divergence is too quick to dismiss models which put zero probability on observed outcomes. While these models are clearly "wrong," i.e. provably not reporting the true distribution of observations, this is usually not our main concern ("all models are wrong, some are useful"). Often we want to compare against model-free baselines (e.g. observed frequencies) which do report zero probability on observed outcomes, and adding parameters to hedge their reports is undesirable. When this is not an issue, KL-divergence is desirable due to its popularity and connections to information theory.

Other proper scoring rules exist, and their divergence functions can be used in model comparisons just as for the logarithmic and quadratic scoring rules.

For example, cosine similarity corresponds to the spherical scoring rule. In general, any Bregman divergence is a feasible way of comparing probabilistic models (i.e. has an associated proper scoring rule).

5 Alternatives

One popular method of scoring probabilistic models is *perplexity*, which is simply an exponentiated version of (8). This exponentiation rewards slight over-reporting of high probability events, but the effect diminishes rapidly with increasing dataset size. Nonetheless, it is theoretically preferable to use the un-exponentiated version for model comparisons.

There are many popular ways of scoring non-probabilistic predictions based on classification accuracy, precision and recall, and so on. These methods can be applied to probabilistic models, for example by ranking outcomes by their reported probability. However, such procedures discard much of the information probabilistic predictions provide, and so are generally less desirable when choosing between probabilistic models.

References

Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102 (477):359–378, 2007.